Specification

Modular Scalable Switching Networks

Field of the Invention

The present invention generally relates to interconnection networks, and in particular to a scalable switching fabric architecture allowing for the building of large switching networks from a common module.

BACKGROUND OF THE INVENTION

Fig. 1 shows the structure of a crossbar matrix 1 in which input lines 10 and output lines 20 are shown to be situated perpendicular to each other and are further shown to be connected at respective crosspoints. The crossbar matrix 1 is non-blocking - any input line may be connected to any output line without blocking other input – to- output connections. For a N x N square crossbar wherein the number of input lines and output lines are the same, (N representing the number of input lines and the number of output lines) the complexity grows by N^2. In addition to complexity in monolithic IC implementation, the package pin constraints will limit the number of input and output lines (N).

Multi-stage Interconnection Networks (MIN) enable building large switches from smaller switches in a structured manner. A three stage (stages 120, 130 and 140) network is shown in Figure 2. The network of Figure 2 is a class of networks based on the Clos network. This type of network belongs to so-called constant stage networks or limited stage networks. In Fig. 2, a Clos network 100 is shown with N input lines 110 and N output lines 150. The N input lines 110 are divided into m groups 111, 112, ..., 119,in Fig.2 we are showing 'n' rather than 'm', each group consisting of n inputs (N = m * n). Each group of n inputs is connected to a (n x k) switch 121, 122, ... 129. The first stage 120 (also called input stage) consists of m groups of (n x k) switches 121, 122, ..., 129, each having n input lines (not shown) and k output lines (not shown). The second stage 130 (also called middle stage) consists of k switches 131, 132, .., 139 of size (m x m). The third stage 140 (also called the output stage) consists of m (k x n) switches 141, ..., 149. The N output lines 150 include the outputs 151, 152, .., 159 of the m switches 141, 142, ..., 149, respectively, in the output stage 140.

With reference to Fig. 2, the outputs 161, 162, .., 169 of the first stage 120 each consist of k lines which will be denoted by $O1_{ij}$ ($1 \leq i \leq m$, $1 \leq j \leq k$), the i index identifies the

switch in the 1st stage and the index j identifies one of the k output lines of the 1st stage of the switch. The input lines 171, 172, .. 179 of the second stage 130 each consist of m lines denoted by $I2_{ij}$ ($1 \le i \le k$, $1 \le j \le m$) the index i identifies one of the k switches and the index j identifies one of the m input lines of the switch. The output lines 181, 182, .. 189 of the second stage 130 each consist of m lines denoted by $O2_{ij}$ ($1 \le i \le k$, $1 \le j \le m$) the index i identifies one of the k switches and the index j identifies one of the m output lines. The input lines 191, 192, .. 199 of the third stage 140 each consist of k lines denoted by $I3_{ij}$ ($1 \le i \le m$, $1 \le j \le k$) the index i identifies one of the m switches and the index j identifies one of the k input lines of the switch. The output lines 151, 152, .. 159 of the third stage 140 each consist of n lines denoted by $O3_{ij}$ ($1 \le i \le m$, $1 \le j \le n$) the index i identifies one of the m switches and the index j identifies one of the n outputs of the switch. The interconnection between stages is as follows:

Output j of switch i in the first stage (O1ij) is connected to input i of switch j in the second stage (I2ji).

Output j of switch i in the second stage (O2ij) is connected to input i of switch j in the third stage (I3ji).

The Clos network 100 of Fig. 2 is denoted as a v(k, n, m). To avoid blocking problems, the design parameters of the network must be selected properly. Clos has shown that the three stage network of Fig. 2 is strictly non-blocking if $k \ge (2n - 1)$. The complexity of switch can be reduced if existing connections can be broken and remade without loss of data in order to establish additional new connections. This type of switch is called a Rearrangeably Nonblocking Switch. The three stage network of Figure 2 is rearrangeably nonblocking if $k \ge n$.

The input stage 120 and output stages 140 of a Clos network consist of m switches. The middle stage 130 of network 100 consists of k switches. The prior art suggest systems with integrated input and output stages and separate middle stage. The problem with prior art is that the structure is not modular, that is the v(k, n, m) network of Fig. 2 can not be constructed from m identical modules.

Therefore, there is a need for constructing a Clos network v(k, n, m) from m identical modules, furthermore there is a need for constructing a modular and scalable Clos networks v(k, n, m), wherein network of different sizes (different values of m) can be constructed from m identical modules and allows building larger networks from a module by adding such

modules as needed. The advantage of a modular structure is that it allows integration into a single module or monolithic Integrated Circuit (IC). The advantages of modular and scalable structure is that it allows building networks of different sizes from the same module or IC.

SUMMARY OF THE INVENTION

Briefly, an embodiment of the present invention includes an expandable network comprising of modules having switches wherein the modules are identical. Furthermore, a method for building a network with varying sizes (different values of m) from a common module is disclosed.

These and other features and advantages of the present invention will become well understood upon examining the figures and reading the following detailed description of the invention.

IN THE DRAWINGS

Fig. 1 shows a switch based on crossbar matrix

Fig. 2 shows an example of Clos network of prior art.

Figs. 3 illustrates a Clos network in accordance with an embodiment of the present invention.

Figs. 4 illustrates a Clos network in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring now to Fig. 3, a Clos network v(k, n, m) 200 is shown in accordance with an embodiment of the present invention. The network 200 includes N input lines 210 and N output lines 250. The N input lines 210 are divided into m groups 211, 212, ..., 219, each group consisting of n input lines (N = m * n). Each group of n input lines is connected to a (n x k) switch 221, 222, ... 229. The first stage 220 (also called input stage) consists of m groups of (n x k) switches 221, 222, ..., 229, each having n inputs 211, 212, ..., 219 and k outputs 261, 262, .., 269. The second stage 230 (also called middle stage) consists of m switches 231, 232, ..., 239 of size (k' x k') each having k' inputs 271, 272, .., 279 and k' outputs 281, 282, ..., 289. The third stage 240 (also called the output stage) consists of m (k x n) switches 241, ..., 249 each having k inputs 291, 292, .., 299 and n outputs 251, 252, ...,259. The N output lines 250 consist of the outputs 251, 252, .., 259 of the m switches 241, 242, .., 249 in the output stage 240.

The input stage 220, the middle stage 230, and output stages 240 of Clos network 200 of the present invention all consist of m switches. The switches 221, 231, and 241 are grouped together and are included in the module 201. In a similar manner, the other modules 202, ..., 209 include the grouping of switches (222, 232, 242), ..., (229, 239, 249), respectively. The network 200 is built from identical modules 201, 202, ..., 209 where each module includes groups of switches (221, 231, 241), (222, 231, 242), ..., (229, 239, 249). Each module includes three switches an input switch of size (n x k), a middle switch of size (k' x k') and an output switch of size (k x n)

A discussion is now presented regarding the selection of the parameter k'. With reference to Fig. 2, the second stage 130 of network 100 consists of k (m x m) switches. The second stage 230 of network 200 consists of m (k' x k') switches 231, 232, ..., 239.

With each (k' x k') switch, a network of q (m x m) switches, can be built wherein q is the quotient of dividing k' by m (q = Q(k'/m) where Q (x/y) denotes the quotient of x divided by y). The m*q inputs and outputs of (k' x k') switch are used and the (k' – m*q) remaining inputs and outputs are unused. The minimum number of equivalent (m x m) switches that is required in the second stage is k. Since in the second stage of network 200, there are m (k' x k') switches, a network of m*q (m x m) switches can be built, therefore parameter k' must be selected such that:

$$\mathbf{m * Q (k'/m) \geq k} \qquad \textit{Eq. 1}$$

The connectivity of the middle stage 230 of the network 200 is now described. A simple way of describing the connectivity of the middle stage 230 is in terms of an equivalent virtual (m x m) switch. That is, the middle stage 230 is first transformed, into an equivalent virtual k (m x m) switches and the connectivity of the equivalent virtual k (m x m) switches is the same as that described previously. There are m (k' x k') switches 231, 232, ..., 239 and as described above, with each (k' x k') switch, a network of q (m x m) switches is built. Starting with the first switch 231, the q*m input and output lines are assigned to the q equivalent (m x m) switches and any remaining input and output lines of the switch are unused. The virtual (m x m) switches are labeled 1, .., q, the assigned input lines are labeled $I2_{ij}$. The index i identifies one of the virtual switches and the index j identifies one of the m input lines of the virtual switch, the assigned output lines are labeled $O2_{ij}$. The index i identifies one of the virtual switches and the index j identifies one of the m outputs. The same process is repeated for the second switch 232, the virtual (m x m) switches are labeled (q+1), .., 2q, the assigned inputs and outputs are labeled $I2_{ij}$ and $O2_{ij}$ respectively similar to the first switch. This process is

repeated until the last switch 239. With k' satisfying Eq. 1, then at least k equivalent (m x m) virtual switches is constructed. The inputs lines of said virtual switches are labeled $I2_{ij}$ ($1 \le i \le k$, $1 \le j \le m$) where the index i identifies one of the k virtual (m x m) switches and the index j identifies one of the m input lines of the switch. The output lines of the virtual switches are labeled $O2_{ij}$ ($1 \le i \le k$, $1 \le j \le m$). The index i identifies one of the k (m x m) virtual switches and the index j identifies one of the m outputs. The interconnection between stages is as follows:

Output j of switch i in the first stage (O1ij) is connected to input i of virtual switch j in the second stage (I2ji).

Output j of virtual switch i in the second stage (O2ij) is connected to input i of switch j in the third stage (I3ji).

We have disclosed a v(k, n, m) switching network using m identical modules and a method of building a v(k, n, m) switching network from m identical modules.

Next, methods for building a scalable v(k, n, m) switching network from a common module will be disclosed. The method includes constructing a switching network 200 from common module 201 and selecting parameter k, k' and specifying a subset of integers $\mathcal{M}$ such that networks with values of m belonging to $\mathcal{M}$ ($m \in \mathcal{M}$) can be constructed with the same common module.

Let the set of divisors of k be denoted by $\mathcal{M}_k$ = { m| m divides k}. If k' is selected to be equal to k (k'= k), it is obvious that m*Q(k'/m)=k for all $m \in \mathcal{M}_k$, and *Eq. 1* is satisfied. Please note that the set $\mathcal{M}_k$ at least includes 1, and k. Therefore one method is as follows:

k' = k and $\mathcal{M}$= $\mathcal{M}_k$ = { m| m divides k} ***(A1)***

Let m_1, m_2, .. denote divisors of k other than 1, and k. If k is prime then the set $\{m_i\}$ is empty. With k' and $\mathcal{M}$ according to *A1* then *Eq. 1* is satisfied for all ($m \in \mathcal{M}_k$), and network of size $\{n, (m_1)*n, (m_2)*n, \ldots, (k)*n\}$can be constructed from identical modules 201 as described previously.

Another method is as follows:

$k' = k = a^s$, and $\mathcal{M}$= { m| $m = a^r$ ($1 \le r \le s$)} ***(A2)***

With k' and $\mathcal{M}$ according to *A2* then *Eq. 1* is satisfied for all ($m \in \mathcal{M}$), and network of size $\{n, (a^1)*n, (a^2)*n, \ldots, (a^s)*n\}$can be constructed from identical modules 201 as described previously. In practice, integer a is typically two (2) and $k= 2^s$, and $m = 2^r$ ($1 \le r \le s$). For example, select $k' = k = 32 = 2^5$. Then networks of size n, $(2^1)*n$, $(2^2)*n$, $(2^3)*n$,

(2^4)*n and (2^5)*n can be constructed from identical modules 201 as described in the embodiment of present invention.

The above method is an exponentially scalable solution. The exponentially scalable method is perfectly acceptable method in certain applications. We now present a linearly scalable method.

Another method to satisfy *Eq. 1* is as follows:

$$k' = k + (M-1),\ \mathcal{M} = \{ m \mid 1 \le m \le M \}, \qquad (A3)$$

With k' and $\mathcal{M}$ according to *A2* then *Eq. 1* is satisfied for all ($m \in \mathcal{M}$). To demonstrate this let k =q*m + r ($0 \le r \le m-1$), and let k'= k + (M-1),

$q*m \le k < (q+1)*m$

If k is divisible by m then r =0, and $m*Q(k'/m) = k$ and *Eq. 1* is satisfied.

If k is not divisible by m then $1 \le r \le (m-1)$, and $(k + (m-1)) \ge (q+1)*m$,

Since $m \le M$ therefore $k' = k + (M-1) \ge (q+1)*m$, and $Q(k'/m) \ge (q+1)$, therefore

$m*Q(k'/m) \ge (q+1)*m > k$ and *Eq. 1* is satisfied.

With k' according to *A3* then you can build a network of size n, 2n, 3n , 4n, ..., M*n, that is you can build a linearly scalable Clos network upto M*n from identical modules 201 as described in the embodiment of present invention.

Several methods for constructing a scalable modular Clos network has been disclosed. Although the method has been described in terms of specific values for {k, k', $\mathcal{M}$}, different values of {k, k', $\mathcal{M}$} that satisfy *Eq. 1* are considered to fall within the true spirit and scope of the invention.

In the method described, some inputs and outputs, specifically (k' – m*q) where q = Q(k'/m), of the 2nd stage of module 201 will be unused since there is not enough inputs and outputs to make an m x m switch. In another embodiment, the parameter k' is selected to be equal to k. After assigning the m*q inputs and outputs, the remaining (k – m*q) inputs of the switch and the first ((m+1)*q – k) inputs and outputs of the next switch are assigned to next virtual (m x m) switch, the remaining inputs and output of the next switch are grouped into groups of m inputs and outputs, forming complete virtual (m x m) switches. With remaining inputs and outputs, the same process is repeated. Since there is no unused inputs of outputs, k'=k, however in this system, some of the virtual (m x m) switches are split between two modules. In order to implement the split virtual (m x m) switches in the middle switch, additional inputs and outputs between adjacent modules is included. Specifically, there is m bi-directional (or 2m unidirectional) signals between adjacent modules to allow operation of

split switches. Fig. 4 shows an embodiment 300 according to the above. The common module 302 includes a 1st (n x k) switch, a 2nd (k, k) switch, a 3rd (k x n) switch, and a first M bi-directional lines 401 and a 2nd M bi-directional lines 402 between adjacent modules for implementing split switches. It should be noted that for the first and last modules, one of the M-bidirectional lines will be unused. A Clos networks v(k, n, m) of size n, 2n, 3n , 4n, ..., M*n, can be constructed from module 302 as described above. Although the above embodiment has been described in terms of a specific method to implement split switches between adjacent modules other methods of implementing split switches between adjacent modules fall within the true spirit and scope of the invention.

Although the present invention has been described in terms of specific embodiments it is anticipated that alterations and modifications thereof will no doubt become apparent to those skilled in the art. It is therefore intended that the following claims be interpreted as covering all such alterations and modifications as fall within the true spirit and scope of the invention.

What I claim is: